云算力新闻大数据平台研究

摘 要：各新闻大数据公司的数据处理能力不同、专业领域不同，本平台研究如何利用各家能力，以低成本的方式实现功能更强大的、数据更全面、分析更准确、速度更快的新闻大数据平台。

关键词：云算力；新闻；大数据；云计算

中图分类号：TP311　　文献标识码：A

文章编号：1671-0134（2019）09-116-02　　DOI：10.19483/j.cnki.11-4653/n.2019.09.034

文 / 徐志强[1]　张守先[2]　李满江[1]

引言

百度、腾讯、新浪、网易、搜狐、今日头条等各公司的许多产品提供了新闻的舆情、热点、热搜、快讯、头条、排行等信息，但是每个公司自身的数据来源不同，所以各自的产品服务各有自己的侧重，比如新浪的微博海量数据的优势，其舆情产品舆情通的热点、影响力等基于微博的分析更有权威性。相比，腾讯的腾讯大数据，基于微信公众号等海量新闻数据的分析更为准确。而综合的新闻大数据平台，数据越全面越好，涉及到网站、微博、微信、APP、公众号、论坛等各类新闻渠道，依赖单一数据，数据单薄、分析结果不全面、不准确。数据越全面，分析结果会越准确、及时、可靠、客观。本平台研究如何利用各大数据公司的能力，以单一大数据平台提供服务，通过整合、细分采购、个性抓取等组合情况下，大大降低各新闻单位、政府部门采购大数据舆情服务的成本，并得到更佳的服务。

1. 何为云算力

大数据的分析，设计到数据量非常大，需要的计算机处理能力要很强，才能短时间内得到想要的结果。算力，主要指计算能力，如比特币矿机的算力（也称哈希率）是比特币网络处理能力的度量单位，即为计算机（CPU）计算哈希函数输出的速度。本文“云算力”一词，借用“算力”一词，来表达整合各家公司独立的云计算能力到一起，从而形成的基于云端的数据处理能力。

2. 平台研究

本平台研究基于各家新闻大数据公司计算能力之上的综合平台，智能化利用各大公司的算力，为新闻单位、政府部门等需要舆情服务的单位，使用单一云平台，就可以发起针对各大大数据公司的数据请求，得到经过智能分析、过滤、排重、本地化、定向处理后的最理想的数据结果，形成最全面、能力最强、最专业、最及时的舆情服务，提供最新新闻、滚动头条、地域新闻、传播榜单、传播路径、趋势分析、热门话题、民生热点、舆情分析、热搜、热词、地域排行、地域热点、热门人物等等不同侧面、角度、地域、领域、群体等分析，涵盖新闻网站类、政府网站类、搜索门户类、论坛社区类、微博类、微信类、新闻客户端类等各类数据源。

2.1 可行性

技术上，本平台依赖于其他公司的处理能力，能选择接入平台的大数据公司，要求有开放接口 API，或有数据推送方式。否则就需要自己抓取结果网页后入库。商务上，购买各家公司的大数据服务，并没有在授权上限制在单一平台多次展示给不同的商业用户，从而可以从各家够买数据服务后，综合整理后展示给通过不同账号登录到本平台的不同目标客户，通过多次销售，从而均摊从各个公司购买的数据服务的费用，从而达到以低价格购买高质量新闻舆情服务的目标。

2.2 实施方案

平台的实施，不仅仅是数据的整合，还要涉及到其他几个方面：

（1）各新闻大数据公司现有服务内容及对接：每个公司某些现有的服务，不用处理就可以直接展示给用户，也具有权威性，符合用户的需要。

（2）各大数据公司处理能力对接：根据各公司接口方式，完成各公司数据服务的对接，利用 API、网页抓取等方式。

基金项目：本文受潍坊市科学技术发展计划项目（项目编号：2019ZJ1162）项目资助。

（3）大数据公司处理结果的数据清洗、去重、元数据统一等，合并成一致的数据。各公司的数据定义方式不一致，比如基本信息、日期格式、打分取值范围等等，需要统一格式、去掉重复数据，清洗成一致的有效数据。

（4）各方数据处理结果整合：加权综合、本地化、定向处理等。在展示数据时，用到各公司数据，需要对其加权后整合，并且根据用户需要，去掉无关的地域的数据，只保留用户关心的、当地的数据。

（5）综合调度：如在用户对某项服务发起请求，能按照需要由平台后台分别对各大数据公司同时发起服务请求，再把返回的结果整合后展示给用户。

当然还包括个性化的本地数据的抓取、个性化新闻舆情服务、多租户管理、不同使用单位的数据分离等等。

平台功能举例：如展示某条新闻的传播效果，就要涵盖报纸、网站、微信、微博、APP、论坛、社交网络等多个渠道的传播数据，才能更全面展示一条新闻的影响力、爆发点、时间线等，新闻受众在不同新闻传播渠道上的比重不同、各新闻渠道时效不同，新闻传播表现在不同新闻传播渠道的爆发期、发散期、削弱期、终结期时间段也各异，需要从各个数据平台抓取结果后，通过加权整合形成一个相对完整客观的时间线曲线，同时保存各新闻渠道的时间线供用户参考，各个渠道的点击量、评论数、受众群体的画像等也需要整合，展示成图、表、曲线等方式，提供给用户。



结语

云算力新闻大数据平台在投入上，比各大数据公司的数据分析平台肯定要少很多，数据存储、运算能力等方面，要求都很低，整合出来的理想效果并不差，并且比单一大数据公司的服务还有加强。当然这些依赖于最终用户数量的多少，来分摊各大数据公司的服务费用，并且运营好这样一个平台，开发、维护的工作量也占一定的比例。媒

参考文献

[1] 孙启虎 . 大数据时代新闻媒体生产和传播策略研究 [J]. 山东农业工程学院学报，2019（2）.

[2] 孙燕，李奎尚，孙建才 . 利用协同空间完成跨报社的重大事件报道 [J]. 中国传媒科技，2018（2）.

[3] 理志强 . 新闻现场报道指挥系统经验谈 [J]. 中国传媒科技，2017（8）.

（作者单位：（1. 潍坊北大青鸟华光照排有限公司；2. 半岛都市报社）